Numerical simulation of protein folding
Rotation report

Student Name: Darya Stepanenko
Student ID: 1701029
Date of Submission:
Unit: Theory of Quantum matter, Nic Shannon
Acknowledgment of guidance by non-faculty unit member: Rico Pohle

I. AIMS

For a long time it has been a mystery how organic molecules turned into structured form of life. The atomic-scale structures organize themselves into a cell-scale structured machinery due to physical laws [1]. But how does it happen? Is it a reproducible and deterministic or rather a spontaneous process? Actually, protein folding is a complex non-equlibrium process and additional help from other molecules, such as chaperons, are needed.

The purpose of this project is to explore a non-equilibrium problem as the protein structure formation from the equilibrium point of view through amino acid interactions [1], using a simplified 'Bike chain' model and numerical Monte Carlo simulations. We would like to investigate:

1. If a chain of charged particles, such as amino acids, can create a stable structure, as seen in proteins?
2. Can the system of a protein be frustrated with complex particle interactions (e.g. long-range interactions)?

II. BACKGROUND

Proteins are large molecules composed of one or more chains of amino acids in a specific order. Amino acids in a protein are joined together by covalent bonding [2] . The chain is called primary structure. Stabilized mostly by hydrogen bonds, the chain folds into alpha helicase and beta sheets, as we call the secondary structure. These helicases and sheets together fold into a spatial structure, interacting with each other, creating a 3D folding, that is called tertiary structure. If a protein is formed by more then one tertiary structured chain, than it is called quaternary structure. Among the interactions, that help folding into tertiary form to happen are electrostatic, Van Der Waals and hydrogen-bond interactions [3]. Interactions are treated differently depending on the distance between amino acids. For amino acids, located close to each others, covalent bonding and bond twisting due to bond order (e.g., double bonds) have a strong influence. On the other hand, for amino acids, located far from each other long-range interactions as Van der Waals and electrostatic interactions play an important role [4].

To tackle the problem of self-organizing organic molecules such numerical method as Monte Carlo (MC) [5] simulations are appropriate, that are rather common in other fields of science, e.g. condensed matter physics. In the following, we will use a simplified (so called 'Bike chain' model) and Monte Carlo simulations to explore the influence of interactions, stated above, on the process of folding.

III. THE BIKE CHAIN MODEL

Protein folding is a complex problem, where many different types of interactions play role. To have a glimpse on this tricky problem, we simplify it to a 2-dimensional case, where we don't consider bond twisting.

We should clarify the statements, what we mean by a chain and by folding in terms of physical and numerical terms. The folding itself is an equilibrium process of conformational changes, that could be considered in terms of a statistical ensemble. We consider a chain as a system with a fixed number of interacting particles, connected by covalent bonds, in contact with a heat bath of temperature T. The chain is described in terms of radius vectors and charges in a 'bike-chain' model, that is explained below.

Covalent bonds between nearest neighbors: amino acid$_i$ and amino acid$_{i+1}$ [see Fig.1a], could be approximated as a fixed rigid rod between particles [see Fig.1b] : particle i (radius vector $\vec{r}_i$ and charge q_i) and particle i+1 (radius vector $\vec{r}_{i+1}$ and charge q_{i+1}). But due to geometry of electron orbitals involved in this covalent bonding, overlapping of next-nearest neighboring amino acids occur: amino acid$_{i-1}$ ($\vec{r}_{i-1}$, q_{i-1}) and amino acid$_{i+1}$ ($\vec{r}_{i+1}$,

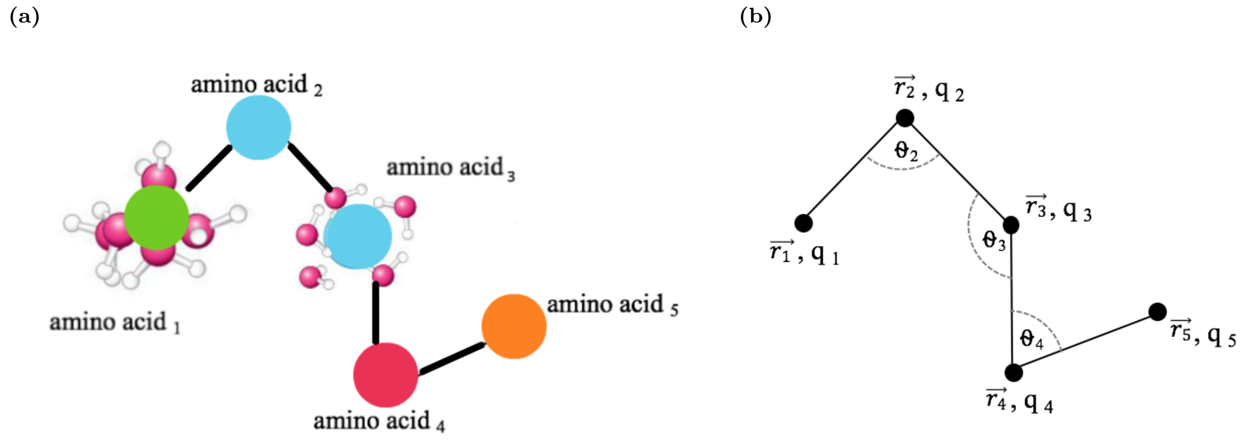


FIG. 1: Simplification of amino acid sequence as a 'bike-chain' of links of fixed length. **(a)** Simplification of amino acids' sequence as a connected charged particles, where different colored circles represent different amino acids, white-blue molecules around amino acid$_1$ and amino acid$_3$ represent water molecules (H_2O) as a solvent **(b)** Bike chain' model. $\vec{r}_i$ is particle's position, θ_i is the angle between $\vec{r}_{i-1}$ and $\vec{r}_{i+1}$, $\Delta\vec{r}(\theta_i)$ is the distance between particle $i-1$ and particle $i+1$.

q_{i+1}), causing the fluctuations of them around some stable configuration. This fluctuations could be described by a spring between two rods connected with each other. The simplified way to describe the amplitude of this overlap can be achieved by using spring with spring constant k between two rods connected with each other.

Long-range interactions we describe with electrostatic Coulomb potential between amino acid$_i$ ($\vec{r}_i$, q_i) and all other amino acid$_j$ ($\vec{r}_j$, q_j), $j \neq i$, that depends on the distance between amino acids and solvent permittivity ε. Summarizing all that we stated above we construct the Hamiltonian for the system:

$$H = \sum_{i,j} U(\vec{r}_i;\vec{r}_j) = \sum_i k(\theta(\vec{r}_i) - \theta_{eq}(\vec{r}_i))^2 + \sum_{i,j} \frac{q_i q_j}{\varepsilon|\vec{r}_i - \vec{r}_j|}, \tag{1}$$

where i is a particle's number at a position $\vec{r}_i$, q_i is a particle's charge.

In real proteins each amino acid in a chain is characterized by the value of negative, positive or neutral charge. In this simplified model all particles are considered to be negatively charged. But the same simulation could be reproduced with any distribution of charges.

So the total energy of the system Eq.(1) is described through the interaction potential $U(\vec{r}_i;\vec{r}_j)$ between particles. The first term in the Eq.(1) is a spring energy as a function of distance between particle $\vec{r}_{i-1}$ and particle $\vec{r}_{i+1}$ with a covalent bond strength k, and long-range interaction is a Coulomb potential with ε solvent permittivity , that is a second term.

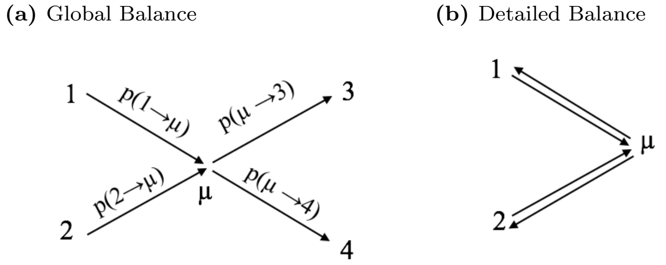
Thus we characterize a chain as covalent bonded particles with bond strength k, and chain folding as a canonical ensemble in a heat bath of temperature T and permittivity ε. To simulate the process of folding for a primary chain into a tertiary structure we minimize the Hamiltonian, stated in Eq.(1), by using numerical Monte Carlo simulations.

IV. METHOD

Monte Carlo simulations

The general idea behind the Monte Carlo method is to provide a sampling of the configurational phase space to obtain statistical quantities of some system's parameters.

We consider a chain of particles, as explained in Sec III, to be in a thermal equilibrium at temperature T. The system is characterized by a discrete number of microscopic chain's configurations μ. E_μ is the energy of the configuration μ. For a system in thermodynamic equilibrium, the probability of finding the state μ corresponds to the Boltzmann weight $p(\mu)$:


(a) Global Balance

(b) Detailed Balance

FIG. 2: Transition rate. **(a)**the condition of global balance states that the rate at which the system makes transitions to and out of a state μ are equal. **(b)** the condition of detailed balance tells that the system should go from μ to ν as often as it goes from ν to μ

$$p_\mu = \frac{1}{Z} e^{-\beta E_\mu} \ , \tag{2}$$

$$Z = \sum_\mu e^{-\beta E_\mu}, \tag{3}$$

where E_μ is the energy of a system in state μ, Z is a partition function, $\beta = 1/k_B T$ and k_B the Boltzmann constant.

Sampling should satisfy following propeties:

1. Markov chain, that means, each iteration should have no memory along the sampling, iteration shouldn't know anything about the previous set of iterations;
2. Ergodicity, that means every possible configuration is allowed;
3. Detailed balance.

The time evolution of a state is coupled to the transition probability $P(\mu \to \nu)$ and can be described by the occupation probability $\omega_\mu(t)$, the measure of how likely it is to find the system in a microstate μ at a time t as shown by the Master equation:

$$\frac{d\omega_\mu}{dt} = \sum_i (\omega_\nu(t) P(\nu \to \mu) - \omega_\mu(t) P(\mu \to \nu)) \tag{4}$$

If the system is in equilibrium then the occupation probability is not time dependent anymore $\frac{d\omega_\mu}{dt} = 0$. So from the Eq. (4) derive that the sum of incoming and outcoming transitions need to be equal [see Fig.2a]. Detailed balance defines the condition that the individual incoming transition is equal to their opposed outcoming transition. It means that system makes transition to and out of the state μ with equal rate, as shown in Fig.2b.

As was stated above, while in equilibrium, the probability of the system to occupy a certain state μ corresponds to the Boltzmann distribution $p(\mu)$. By introducing the distribution Eq.(3) and detailed balance condition into Eq.(4) we obtain:

$$\frac{p_\nu}{p_\mu} = \frac{P(\mu \to \nu)}{P(\nu \to \mu)} = e^{-\beta(E_\nu - E_\mu)} \ , \tag{5}$$

where the ratio of transition probabilities is written by the Boltzmann weight of the energy differences between states.

Metropolis suggested [6] that, if a new state has an energy, lower than the present one, we always accept the transition to that new state. If the new state has a higher energy, then we accept it with the probability Eq.(6) as shown in Fig.3

$$P(\mu \to \nu) = 1 \ , \ if \ E_\nu < E_\mu \tag{6a}$$

$$P(\mu \to \nu) = e^{-\beta(E_\nu - E_\mu)} \ , \ if \ E_\nu >= E_\mu \tag{6b}$$

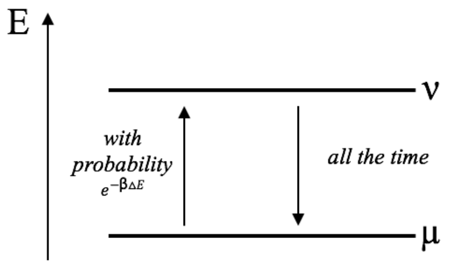


FIG. 3: Energy scheme for the transition probability between state μ and ν in the Metropolis argument [Eq.(6)]

Monte Carlo simulation on the computer

On the computer the algorithm, described above, works as following. The program chooses one random particle r_i in a chain and changes locally the angle θ_i between rods attached to this particle. The state with a new angle is accepted if this change is energetically favorable. Repeat this step for N randomly chosen particles, where $N =$ number of angles in a chain. This process is called Monte Carlo step.

The evaluation of probability in the Metropolis argument in Eq.(6) takes place with a random number $R_0 \in [0, 1]$. If the random number R_0 is smaller than the Boltzmann weight, that corresponds to the energy difference between state μ and ν, then we accept the transition. How does the most deterministic machine such as a computer provide a random number is another question. Roughly speaking there is a piece of computer code that deterministically produces statistically uncorrelated numbers. Randomness of programs could be different. What is random for one application may not be random enough for another. The length of the sequence is usually related to the periodicity of random generator. In this work we use python library numpy.random, that allows to get uncorrelated statistical sampling in Monte Carlo simulations.

Measurements

Thermodynamic parameters are connected to statistical mechanics variables. Monte Carlo simulations provide a sampling of configurational phase space of the system to obtain statistical quantities of system parameters. So we can connect statistical mechanics - microstates with thermodynamics observables.

To connect statistical and thermodynamic sides of one problem we run the simulation for a suitably long time until it comes to equilibrium at certain temperature T, permittivity ε and covalent bond constant k. This period is called equilibration time. Then we measure the quantity we are interested in (energy and radius of gyration) over suitably long period of time, allowing all possible configuration to occur, and average it. If X_μ is an observable of the system, X_μ could be the energy ($X_\mu = E_\mu$) or any other physical quantities:

$$\langle X \rangle = \frac{1}{N_m} \sum X_i \ , \tag{7}$$

where N_m the number of measurements of X.

We would like to measure how tight and stable the folded chain is. The stability of the system could be characterized by evaluating Hamiltonian Eq.(1), that shows if the system reaches its minimum energy state and stays in it. Another characteristic, we are interested in, is the geometry of the protein to characterize the tightness of folding. The parameter, that tells basically how far the particles are extended from its center of mass, is the radius of gyration:

$$R^2 = \frac{\sum_i (\vec{r}_i - \vec{r}_{c.m})^2}{N}, \ \vec{r}_{c.m} = \frac{\sum_i \vec{r}_i}{N} \tag{8}$$

To obtain statistically relevant quantities, we make measurements of two parameters: energy and radius of gyration, and average it by using Eq.(7).

V. RESULTS

When floating in the solution of some temperature, amino acids experience an influence from it. System parameters as solvent permittivity ε, temperature T and covalent bond strength k have an influence on the stability of the chain and reproducibility of the simulation. While ε and k characterize stabilizing forces, T is in charge of disturbing thermal fluctuations.

So, if we look at the Boltzmann probabilities in Metrorpolis argument Eq.(6), we notice these three parameters. If we divide the denominator and numerator in the Boltzmann exponent by covalent bond strength value k, the exponent remains unchanged. So, we divide $\frac{1}{\varepsilon}$ and T by k and measure them in units of k. So we have two parameters instead of three: $\frac{1}{k\varepsilon}$ and $\frac{T}{k}$. It means that actually two combinations of these three parameters influence the system. To make it more convenient, fix $k = 1$.

Solvent permittivity ε (measured in units of k) is an interesting parameter. It mimics the role of a solvent, increasing and decreasing screening effects on the particles' charges. In both the presence and the absence of the solvent radius of gyration fluctuates as shown at Fig.4b and Fig.4e. But it is worth to notice that fluctuations happens around different means. In the absence of Coulomb the chain is tightly packed, the mean of radius of gyration is low Fig.4b, so the relative error is bigger than in the presence of Coulomb interactions Fig.4e, where the mean is higher.

It could mean in the system with lower ε after significantly many Monte Carlo steps the chain ends up with a relative stable configuration Fig.4d with stable in time radius of gyration Fig.4e.

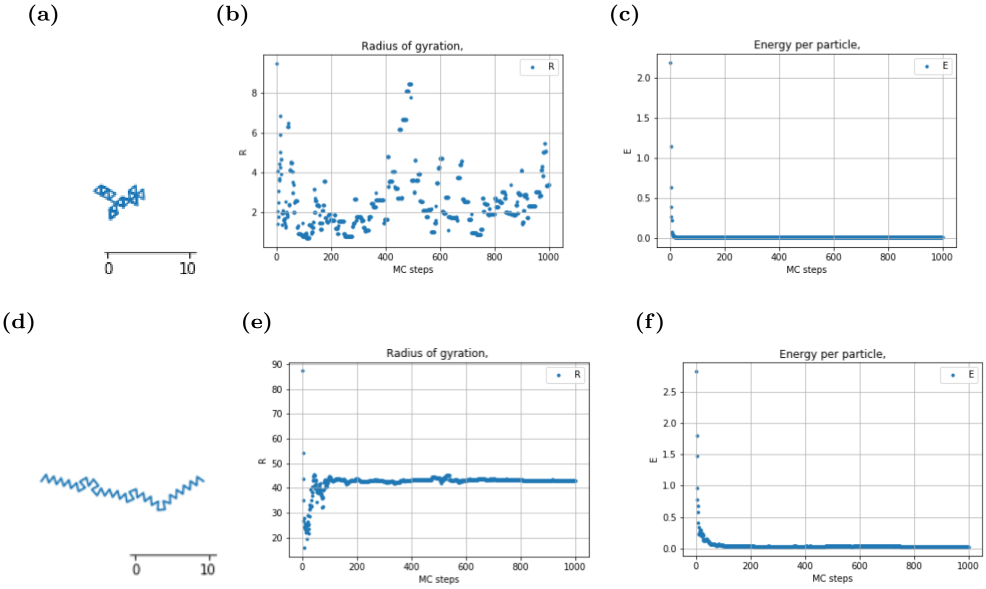


FIG. 4: **Upper row:** The system with high solvent screening;**(a)** Configuration of a chain after 999MC steps;**(b)**The evolution of radius of gyration with MC steps; **(c)**The evolution of internal energy per particle with MC steps. **Lower row:** The system with low solvent screening, $\varepsilon k = 20$;**(d)**Configuration of a chain after 999MC steps;**(e)**The evolution of radius of gyration with MC steps;**(f)**The evolution of internal energy per particle with MC steps. System's parameters: number of particles = 50, $k = 1$, $\frac{T}{k} = 0.0001$, $\theta = 60\circ$

By keeping k and T the same but changing the solvent permittivity, we change only the Coulomb interactions between the particles in a chain and moving back and forth between two regimes: strong and weak electrostatic interactions. In the $\lim_{\varepsilon\to\infty}$ the situation with only local, covalent bonded interactions appears, leading to the decrease of the radius of gyration as shown at Fig.5b. In the $\lim_{\varepsilon\to 0}$ covalent bond influence becomes negligible and the Coulomb interactions takes the leading position at the same time increasing the interaction potential as shown at Fig.5a.

The influence of such parameter as temperature (measured in units of k) helps to escape from local minima. This would work only if the local minimum barrier is compatible with thermal fluctuations. But at the same time the increase of temperature started the battle between covalent bonds together with Coulomb interactions versus thermal fluctuations. At low temperatures covalent bonds and Coulomb interactions win and what we have is a stable structure in energetically favorable equilibrium state. While with the increase of temperature thermal fluctuations win, and the stability vanishes, the Coulomb energy can't stretch the chain anymore and the folding becomes more compact with low radius of gyration as shown at the Fig.5b but chaotic, that is shown by the increase of energy at Fig.5a.

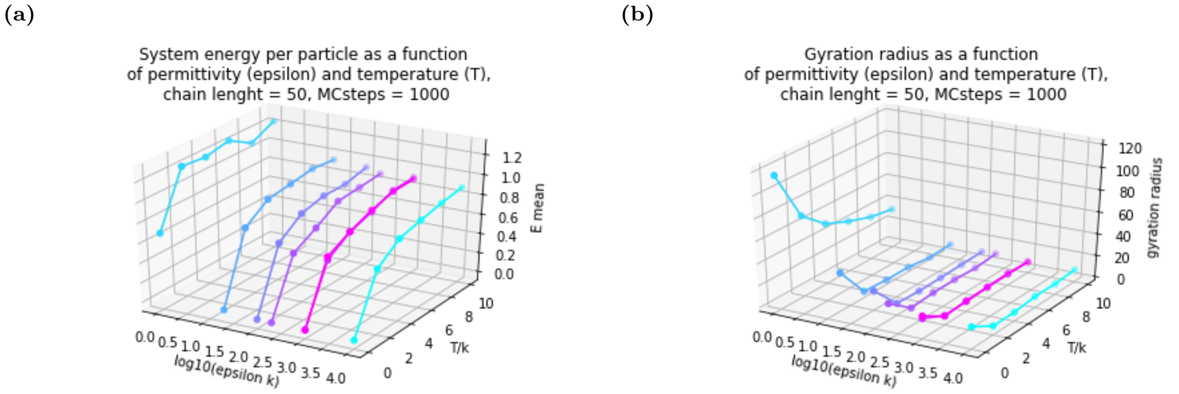


FIG. 5: **(a)** System energy and **(b)** radius of gyration as a function of permittivity εk and temperature T/k. Chain length = 50, all particles are negatively charged, MC steps = 999.

By repeating the same simulation with the same system properties twice, we get the same energy, the same radius of gyration, but the structure conformations are different. We don't get the same stable configuration more than once. It shows that the chain behaves like a glassy system, with lots of similar local minima, and requires additional help from other cell's structures to reach a global minima. But than, it is a non equilibrium problem and a completely different story.

VI. ADDITIONAL PLANS

- The Monte Carlo algorithm samples over an ensemble and chooses the most energetically favorable one. There is a very large number of freedom in an unfolded chain. The molecule has a huge number of possible configurations to try before finding the lowest energy state. That process requires long time for a random search. In order to shorten this time, we would like to introduce elongation idea, where the folding starts with short chain and elongates by adding new amino acid$_i$ $\vec{r}_i$ along with folding process. It is worth to notice that when implementing elongation, we assume that new particle doesn't change the detailed balance statement. But it appears, even with allowing the system to get stabilized for some time before adding a new particle, the time requires for stabilizing the whole chain isn't influenced significantly for short chains as 50 particle length Fig.6. But this idea requires more investigation.

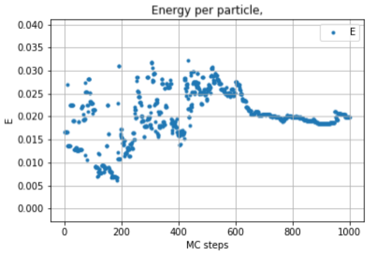


FIG. 6: System energy for the system with 1 amino acid per 10 MC steps elongation. System characteristics: 50 particles negatively charged chain with bonded and non-bonded interactions, elongation step = 1 particle/10steps, $\varepsilon k = 20$

- During our research we studied chains with different length, but the longest was 50 particles long. This length restriction is connected with computational time, required for the simulation. Real proteins consist of a couple of hundred amino acids, that is obviously longer than 50. But we are not sure that scaling works for this problem. Longer chains could show different qualitative behavior.
- It is worth to notice that in this work only negatively charges amino acids were considered. It is worth to play with differently charged and uncharged amino acids.

VII. CONCLUSION

This work was aimed to explore on a simplified model if proteins are reproducible and stable structures without any outside help, for example, as chaperons in a cell. We found out that actually the chain of amino acids could fold into a stable configuration of a protein due to covalent bond interactions and Coulomb interactions. But receiving the same folding configurations at least twice was impossible, even without changing system parameters ε, T, k. It could mean that the system has many local minimum with close energies, that correspond to different configurations. At this work we considered a simplified 2D model of a system, and some forces were omit, for example bond twisting and van der Waals interactions, that could influence the stability of the folded chain. As we know, simplifying the 3D systems to 2D models could lead to the lost of some principle physics, so these forces are worth to be considered in the future.

[1] Alan R. Fersht. From the first protein structures to our current knowledge of protein folding: delights and scepticisms. *Nature Reviews Molecular Cell Biology*, 9:650, 06 2008.

[2] Jin Xiong. *Essential bioinformatics*. Cambridge University Press, 2006.

[3] Michael Levitt, Miriam Hirshberg, Ruth Sharon, and Valerie Daggett. Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Computer Physics Communications*, 91(1):215 – 231, 1995.

[4] Yong Duan, Chun Wu, Shibasish Chowdhury, Mathew C Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, James Caldwell, Junmei Wang, and Peter Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem*, 24(16):1999–2012, Dec 2003.

[5] M. E. J Newman and G. T Barkema. *Monte Carlo methods in statistical physics*. Clarendon Press, Oxford, 1999.

[6] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state by fast computing machines. 21:1087–, 06 1953.